OncoTargets and Therapy

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Identification of the predictive genes for the response of colorectal cancer patients to FOLFOX therapy

This article was published in the following Dove Press journal:
OncoTargets and Therapy

Hengjun Lin
Xueke Qiu
Bo Zhang
Jichao Zhang

Department of Tumor, Anus and Intestine, Jinhua People's Hospital, Jinhua, Zhejiang 321000, China

Background: Colorectal cancer is a malignant tumor with high death rate. Chemotherapy, radiotherapy and surgery are the three common treatments of colorectal cancer. For early colorectal cancer patients, postoperative adjuvant chemotherapy can reduce the risk of recurrence. For advanced colorectal cancer patients, palliative chemotherapy can significantly improve the life quality of patients and prolong survival. FOLFOX is one of the mainstream chemotherapies in colorectal cancer, however, its response rate is only about 50%.

Methods: To systematically investigate why some of the colorectal cancer patients have response to FOLFOX therapy while others do not, we searched all publicly available database and combined three gene expression datasets of colorectal cancer patients with FOLFOX therapy. With advanced minimal redundancy maximal relevance and incremental feature selection method, we identified the biomarker genes.

Results: A Support Vector Machine-based classifier was constructed to predict the response of colorectal cancer patients to FOLFOX therapy. Its accuracy, sensitivity and specificity were 0.854, 0.845 and 0.863, respectively.

Conclusion: The biological analysis of representative biomarker genes suggested that apoptosis and inflammation signaling pathways were essential for the response of colorectal cancer patients to FOLFOX chemotherapy.

Keywords: colorectal cancer, FOLFOX therapy, support vector machine, minimal redundancy maximal relevance, incremental feature selection, chemotherapy response

Introduction

Colorectal cancer is a malignant tumor that seriously endangers people's health. In recent years, the incidence of colorectal cancer has significantly increased and has become the third most common type of cancer. In the past few decades, due to the early detection and treatment, many countries have improved the survival rate of colorectal cancer. Especially in some developed countries, the 5-year survival rate has reached more than 65%.[1]

Treatment options for colorectal cancer include chemotherapy, radiotherapy and surgery.[2] In general, surgical removal of the affected tumor and any adjacent intestines can effectively eliminate cancer cells and reduce the risk of cancer spreading. Chemotherapy also occupies an important role in the treatment of colorectal cancer. Postoperative adjuvant chemotherapy in early colorectal cancer can reduce the risk of recurrence. For patients with advanced colorectal cancer who are inoperable, palliative chemotherapy can significantly improve the life quality of patients and prolong survival.

Generally, the combination of chemotherapeutic agent results in significantly increased response rates and improved survival.[3] Current combination chemotherapy

Correspondence: Hengjun Lin
Department of Tumor, Anus and Intestine, Jinhua People's Hospital, Jinhua, Zhejiang 321000, China
Tel +86 138 5798 8075
Email 13857988075@163.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/OTT.S167656

includes 5-fluorouracil (5-FU)/leucovorin with oxaliplatin (FOLFOX), 5-FU/leucovorin and irinotecan (FOLFIRI), capecitabine and oxaliplatin (CAPEOX/XELOX) and 5-FU/leucovorin/oxaliplatin and irinotecan (FOLFOXIRI).

FOLFOX chemotherapy has proven to be effective in the treatment of unresectable metastatic colorectal cancer.[4] Studies have suggested that patients with stage III colorectal cancer, who receive adjuvant FOLFOX chemotherapy, experience an improved disease-free and overall survival.[5] However, about half of the patients were unable to benefit from the treatment and even suffered from neurotoxicity.[6]

There have been several studies that are trying to predict the FOLFOX chemotherapy response.[7,8] It has been reported that MTHFR germinal polymorphism is a potential strong predictor of response to FOLFOX therapy, and the response rate to FOLFOX increases continuously with the number of favorable MTHFR alleles.[7] Another reported biomarker is SMURF2. It was highly expressed in non-responders for FOLFOX therapy.[8]

To systematically investigate the response mechanisms of FOLFOX chemotherapy in colorectal cancer patients, we collected three gene expression datasets of colorectal cancer patients with FOLFOX therapy and identified the genes that can predict responders to FOLFOX therapy for colorectal cancer using advanced machine learning methods. The biological analysis of several representative signature genes, such as *MLKL*, *CC2D1A*, *LPL*, *PAGE4* and *SLC26A9*, suggested that apoptosis and inflammation signaling pathways were the essential pathways that controlled the response of colorectal cancer patients to FOLFOX chemotherapy.

Methods

The gene expression profiles of colorectal cancer patients with FOLFOX therapy

We searched Gene Expression Omnibus (GEO) database and found three datasets of colorectal cancer patients with FOLFOX therapy.

The gene expression profiles of colorectal cancer patients with FOLFOX therapy were combined from three datasets downloaded from GEO with accession number of GSE19860, GSE28702 and GSE72970. The platform of these three datasets was the same. They all used Affymetrix Human Genome U133 Plus 2.0 Array.

These three datasets were generated by different researchers from different labs. To minimize the systemic bias, the raw CEL files were downloaded and processed together using R package affyPLM and affy.[9] The gene expression levels of probes were quantified with MAS5 method[10] and normalized with quantile method. The probe expression levels were transformed into gene expression levels using R package gahgu133plus2cdf and gahgu133plus2.db. There were 18,733 genes with expression levels that were used as features to predict whether a colorectal cancer patient will respond to FOLFOX therapy.

In GSE72970 dataset, there were 20 colorectal cancer patients with FOLFOX response and 12 colorectal cancer patients without FOLFOX response. In GSE28702, there were 42 colorectal cancer patients with FOLFOX response and 41 colorectal cancer patients without FOLFOX response. In GSE19860, there were nine colorectal cancer patients with FOLFOX response and 20 colorectal cancer patients without FOLFOX response. Together, there were 42 colorectal cancer patients with FOLFOX response who were considered as positive samples and 41 colorectal cancer patients without FOLFOX response who were considered as negative samples. The sizes of positive and negative samples are shown in Table 1. The clinical information of the 144 colorectal cancer patients from GEO is given in Table S1.

Rank the discriminative genes using mRMR method

The minimal redundancy maximal relevance (mRMR) method[11] is widely used to select discriminative features.[12–17] The mRMR software downloaded from http://home.penglab.com/proj/mRMR/ was used to perform the feature ranking.

It works as follows: first, let us represent all the 18,733 genes, the selected m genes and the to-be-selected n genes using Ω, Ω_S and Ω_t, respectively. The relevance I of gene g from Ω_t with FOLFOX response r can be measured with mutual information (I):[18,19]

$$D = I(g, r) \quad (1)$$

The redundancy R of the gene g from Ω with the selected genes in Ω_S are

Table 1 The sizes of positive and negative samples

Dataset number	Number of positive samples[a]	Number of negative samples[b]	Sample size
GSE72970	20	12	32
GSE28702	42	41	83
GSE19860	9	20	29
Combined	71	73	144

Notes: [a]Positive samples: colorectal cancer patients with FOLFOX response. [b]Negative samples: colorectal cancer patients without FOLFOX response.

$$R = \frac{1}{m}\left(\sum_{g_i \in \Omega_s} I(g, g_i)\right) \quad (2)$$

The algorithm tries to find the best gene g_j from Ω_t that has maximum relevance with FOLFOX response r and minimum redundancy with the selected genes in Ω_S by maximizing the function below

$$\max_{g_j \in \Omega_t}\left[I(g_j, r) - \frac{1}{m}\left(\sum_{g_i \in \Omega_s} I(g_j, g_i)\right)\right] (j = 1,2,\ldots,n) \quad (3)$$

After N rounds of evaluation procedure, all the genes from Ω_t will be ranked

$$S = \left\{g'_1, g'_2, \ldots, g'_r, \ldots, g'_N\right\} \quad (4)$$

The mRMR rank represents the discriminating power of the gene.

To reduce the computational time, only the top 500 mRMR genes were analyzed in the following steps.

Identify the predictive genes using incremental feature selection (IFS) method

To evaluate the prediction performance of mRMR genes, IFS method[20–26] was applied to select the genes with greatest prediction power. The IFS method is a wrapped feature selection method that combines the feature selection with classifier construction. We used Support Vector Machine (SVM) as the classifier. To be specific, the SVM function in R package e1071 was used to construct the classifier.

IFS is a process of iteration that adds genes one by one based on the mRMR ranking and then evaluates the classification performance of the selected genes. Each time, the top k genes from the mRMR table were selected and used to build the classifier that predicts whether a colorectal cancer patient will respond to FOLFOX therapy. The performance of each classifier was evaluated with leave-one-out cross validation (LOOCV).

The three major measurements for a classifier, sensitivity (Sn), specificity (Sp) and accuracy (ACC), were calculated.

$$S_n = \frac{TP}{TP + FN} \quad (5)$$

$$S_p = \frac{TN}{TN + FP} \quad (6)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

In these equations, TP, TN, FP and FN stand for true positive samples, true negative samples, false positive samples and false negative samples, respectively.

In this study, the colorectal cancer patients with FOLFOX response and the colorectal cancer patients without FOLFOX response were considered as positive and negative samples, respectively.

After 500 rounds of IFS evaluation, an IFS curve can be plotted. The x-axis was the number of used genes, and the y-axis was the LOOCV accuracy. Based on the IFS, we can easily see how many genes should be used to classify the colorectal cancer patients with FOLFOX response and the colorectal cancer patients without FOLFOX response.

The visualization of how predictive the genes are for FOLFOX response

After we identified the predictive genes using mRMR and IFS methods, we tried to visually investigate how good they can classify the colorectal cancer patients with FOLFOX response and the colorectal cancer patients without FOLFOX response.

Principal component analysis (PCA)[27] was performed to extract the first and second principal component (PC) of the selected genes. PCA is a widely used multivariate statistical method and can capture most of the gene expression variability.[27] With the dimensionality reduction via PCA, the high dimension gene expression profiles can be mapped onto two dimensions of PC1 and PC2, which can explain the most variance observed in the data. Since it is unsupervised, the 2D-PCA plot will give an intuitive view of how close each sample is to each other.

Another method that we applied was two-way hierarchical clustering of both colorectal cancer patients and selected genes. From the heatmap, we can not only explore whether the colorectal cancer patients with FOLFOX response and the colorectal cancer patients without FOLFOX response were clustered into different groups but also know which genes were highly expressed or lowly expressed in the colorectal cancer patients with FOLFOX response.

Results and discussion

The top discriminative genes ranked with mRMR method

The mRMR can rank the genes based on not only their relevance with the FOLFOX responses of colorectal cancer patients but also the redundancy with each other. Therefore, the discriminative genes identified by mRMR methods will be compact, which means the highly co-expressed genes will

not all be selected, only the best representative gene will be chosen. We obtained the top 500 most discriminative genes using the mRMR method. These 500 genes will be further optimized using IFS method.

The predictive genes selected based on IFS method

We used different number of top mRMR genes to construct the SVM classifier. Based on how accurate the model can classify the colorectal cancer patients into the right FOLFOX response groups, we plotted the IFS curve in which the x-axis was the number of genes and the y-axis was the LOOCV accuracy. The IFS curve is shown in Figure 1.

As shown in Figure 1, the peak located at the position of using top 138 genes. Its accuracy was 0.854, which was the highest. We also calculated its sensitivity and specificity, which were 0.845 and 0.863, respectively. The top 138 genes are given in Table S2. The confusion matrix of actual responses and predicted responses is given in Table 2. We calculated the CIs of prediction performance using function sensSpec from R package epibasix[28] and the 95% CIs for sensitivity and specificity were (76.1, 92.9) and (78.4, 94.2), respectively.

Although the performance of 138 genes was best, the accuracy of the top ten genes had already been over 0.8. The sensitivity and specificity for the ten gene classifier were 0.732 and 0.890, respectively. The top ten genes are given in Table 3.

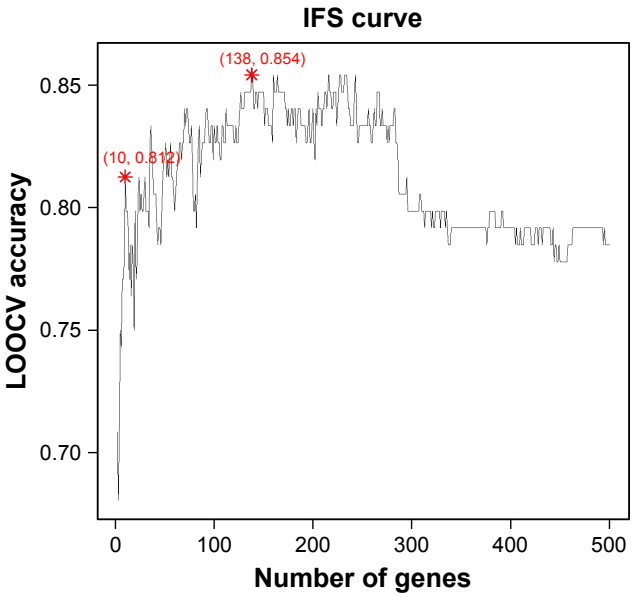

Figure 1 The IFS curve of how the classifiers were based on different number of gene performance.
Notes: The x-axis was the number of genes used to build the classifier and y-axis was the prediction accuracy evaluated with LOOCV. The peak of IFS curve was accuracy of 0.854 when 138 genes were used. But even when only top ten genes were used, the accuracy was over 0.8.
Abbreviations: IFS, incremental feature selection; LOOCV, leave-one-out cross validation.

Table 2 The confusion matrix of actual responses and predicted responses based on 138 genes

Number of patients	Actual responders	Actual non-responders
Predicted responders	60	10
Predicted non-responders	11	63

The first gene was LOC100009676, which was understudied and did not have too much known functions.

The second gene was Lnc-ZNF461, which has been reported to be associated with non-small-cell lung cancer (NSCLC).[29] It was involved in immune response and can promote NSCLC progression by interacting with SLA2, DEFB4A, LAT and LIME1.[29]

The third gene was MLKL, a necroptosis kinase. It was reported that MLKL was involved in immune activation in cancer cells.[30] Chemotherapy kills MLKL$^{-/-}$ cancer cells, and due to MLKL deficiency, the dying cancer cells will not cause immune response. MLKL may function through ICD signaling pathway. A recent publication by Sun et al[31] found that small-molecule analogs of SMAC mimetic in association with MLKL-pDNA and z-VAD-fmk showed antitumor effects in colorectal cancer cells in vitro via induction of RIP3-dependent necroptosis. All these findings have confirmed MLKL as a good chemotherapy response biomarker.

Another interesting gene was CC2D1A, a remarkable member of various signaling pathways, such as nuclear factor kB, PDK1/Akt, cAMP/PKA and Notch. Notch pathway is a well-studied colorectal cancer pathway.[32,33] It has also been reported to be involved in the antiviral pathway by interacting with TBK-1 and IKKε and acts as a transcriptional repressor of serotonin and dopamine receptor genes.[34] CC2D1A silencing can induce apoptosis and increase chemotherapy sensitivity by decreasing Akt kinase activity.[35]

The responders and non-responders were different on the first PC

To intuitively explore the difference of responders and non-responders, we calculated the first and second PCs of the

Table 3 The top ten mRMR genes

Order	Name	Score
1	*LOC100009676*	0.131
2	*ZNF461*	0.101
3	*MLKL*	0.072
4	*MGC15885*	0.083
5	*MBTD1*	0.071
6	*CC2D1A*	0.067
7	*FAM104A*	0.061
8	*KIF3B*	0.060
9	*SYTL1*	0.060
10	*EML6*	0.057

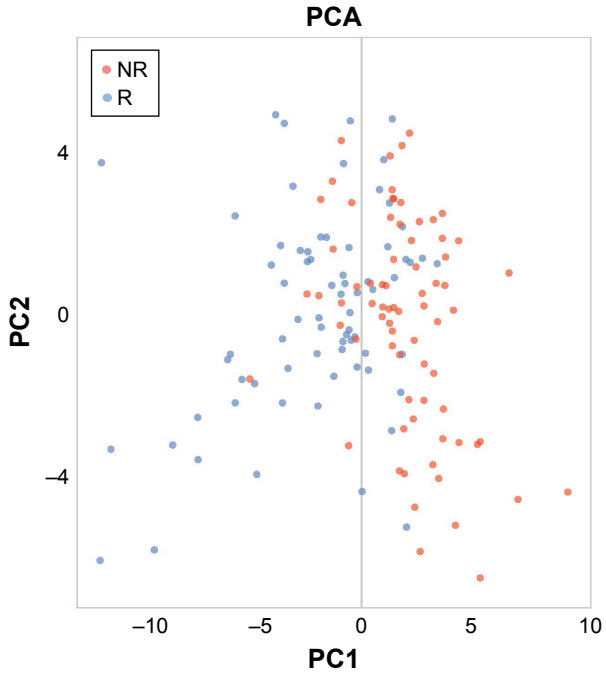

Figure 2 The PCA plot of responders and non-responders.
Notes: The x-axis was the first PC and y-axis was the second PC. The red dots were NR and the blue dots were R. It can be seen that most responders were in area of PC1<0, while most non-responders were in the area of PC1>0. R and NR were different on the first PC.
Abbreviations: PCA, principal component analysis; PC, principal component; NR, non-responders; R, responders.

138 genes and plotted the PCA of responders (blue dots) and non-responders (red dots) in Figure 2. PC1 represented 8.7% variance, while PC2 represented 4.7% variance.

It can be seen that most responders were in area of PC1<0, while most non-responders were in the area of PC1>0. The responders and non-responders were different on the first PC.

The highly expressed genes in FOLFOX responders and non-responders

Although the PCA plot clearly demonstrated the difference of responders and non-responders, we were interested in identifying the highly expressed genes in FOLFOX responders and non-responders, which may reveal the biological mechanisms of FOLFOX response in colorectal cancer. Therefore, we plotted the heatmap of the 138 genes in the responder and non-responder colorectal cancer patients (Figure 3).

It can be seen that the responders and non-responders were clearly clustered into two groups and correspondingly, the 138 genes were also clustered into two groups. The top cluster of genes was highly expressed in responders, and the bottom cluster of genes was highly expressed in non-responders.

We have listed the highly expressed genes in FOLFOX responders whose fold change was greater than 1.5 and the lowly expressed genes in FOLFOX responders whose fold change was smaller than 0.67 in Tables 4 and 5, respectively.

For the highly expressed genes in FOLFOX responders, CRYBB1 was one of the highly mutated genes in microsatellite instability colorectal cancers.[36]

NEUROG3 played important roles in intestinal enteroendocrine cells and was repressed by the growth factor-independent one transcription factor (GFI1) that was normally expressed in Paneth and goblet cells of colon.[37]

LPL is a crucial enzyme for intravascular catabolism of triglyceride-rich lipoproteins. The alteration of LPL may let the cell acquire growth advantage and develop malignancy.[38] The LPL gene deficiency increases cancer risk. The tumor suppressive effects of LPL have been verified in animal models; due to its roles in inflammation, it is a great general target for chemotherapy.[39]

CYP4F is a member of the CYP/CYP450 superfamily of enzymes. It was highly expressed in prostate cancer and RNAi experiments, which suggested that CYP4F was important for cell growth and survival.[40]

PAGE4 is a member of GAGE family, which is highly expressed in various tumors.[41–43] It has been reported that PAGE4 expression can predict liver metastasis of colorectal cancer.[44]

For the lowly expressed genes in FOLFOX responders, SLC26A9 has colon-specific functions, such as transport of glucose, organic acids, metal ions and mineral absorption.[45] Its low expression may affect the growth of tumor cells.

The limitations and potential improvements of this study

Although this study identified candidate genes for chemotherapy response for colorectal cancer and revealed highly possible mechanism, there were several limitations:

Since this was a bioinformatics study, we did not validate our results with biological experiments. This limited the discovery of novel mechanisms. To reduce the effects of lacking experimental validation, we did thoroughgoing literature survey and proposed the possible mechanisms based on confirmed biological functions of candidate genes from published papers.

The sample size of this study was small, even though we collected all publicly available gene expression profiles from the largest gene expression database, GEO. In the next step, we will collect colorectal cancer patients with chemotherapy from our hospital and build a large independent test dataset.

The number of genes was still too large. We will try more advanced feature selection methods to further reduce the

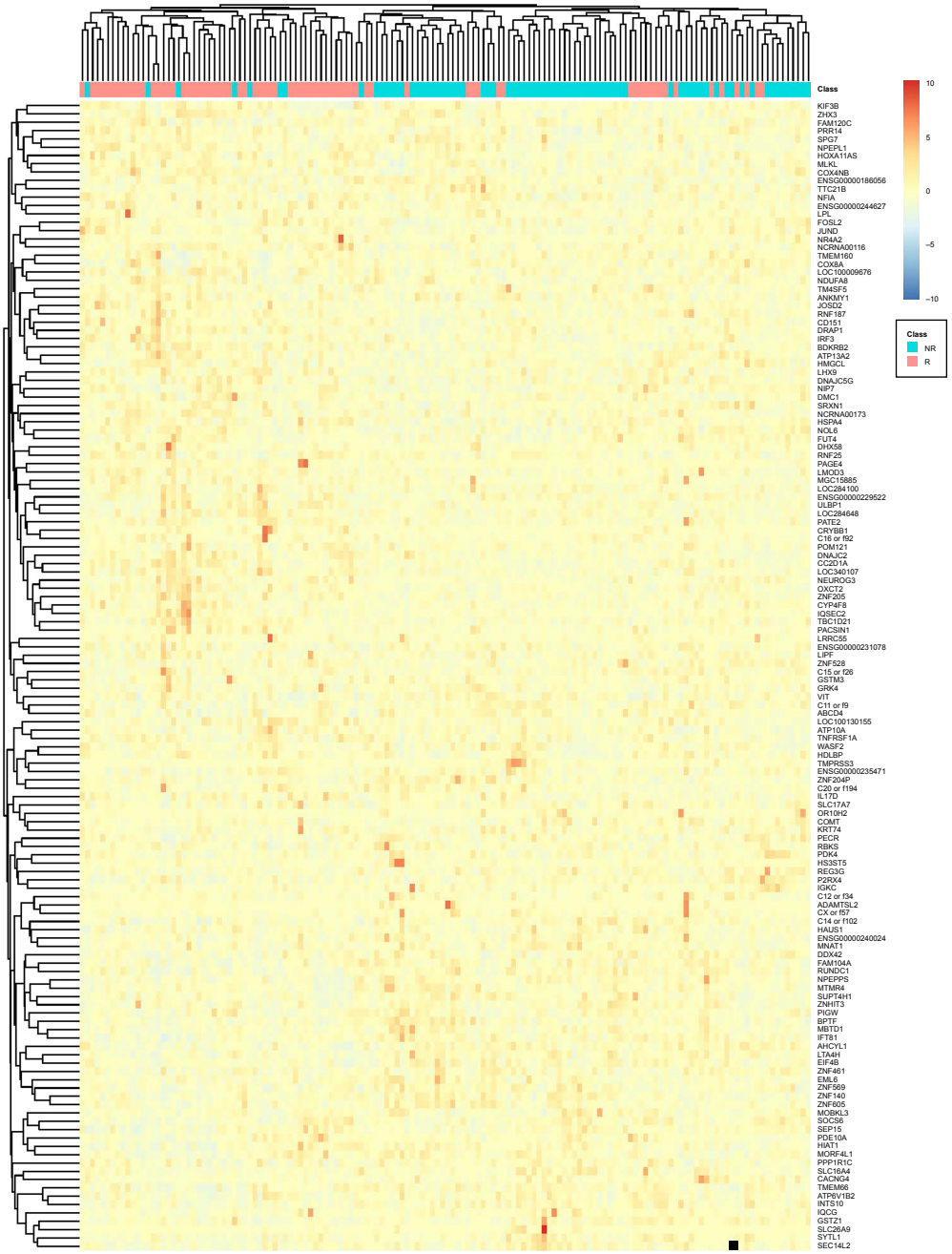

Figure 3 The heatmap of the 138 genes in the responder and non-responder colorectal cancer patients.
Notes: Each row corresponded to the scaled gene expressed level of a gene. The warmer colors indicated higher expression level and the colder colors indicate lower expression levels. Each column corresponded to a colorectal cancer patient who may be responder (red) and non-responder (green) to FOLFOX therapy. It can be seen that the responders and non-responders were clearly clustered into two groups and correspondingly, the 138 genes were also clustered into two groups. The top cluster of genes was highly expressed in responders and the bottom cluster of genes was highly expressed in non-responders.
Abbreviations: NR, non-responders; R, responders.

number of selected genes. The exhaust search strategies can be applied within the 138 genes to find the optimal 3–5 genes.

The clinical information should be documented carefully. Since the data we analyzed were from GEO, much clinical information of the patients was unknown. Analyzing the clinical information may provide novel insight. For example, within the 141 colorectal cancer patients, 117 samples were from primary sites and 27 samples were from metastatic lesions. But, we found that all 27 metastatic samples were predicted with the correct responses, as shown in Table S1 in which the third and sixth columns are actual responses and predicted responses, respectively. There may be two reasons of why the metastatic lesions can predict chemotherapy response: 1) the gene expressions between primary tumors

Table 4 The highly expressed genes in FOLFOX responders

Gene name	Mean in NR[a]	Mean in R[b]	Fold change[c]
MGC15885	11.2	23.8	2.1
ENSG00000244627	15.7	32.8	2.1
CRYBB1	7.6	15.1	2.0
NEUROG3	14.1	26.9	1.9
LOC284100	23.5	43.1	1.8
PACSIN1	9.7	17.2	1.8
LPL	179.4	306.0	1.7
LOC340107	18.0	30.5	1.7
C16orf92	16.4	26.2	1.6
CYP4F8	17.6	27.8	1.6
PAGE4	41.3	64.8	1.6

Notes: [a]NR, colorectal cancer patients without FOLFOX response. [b]R, colorectal cancer patients with FOLFOX response. [c]Fold change, R/NR.

and metastatic lesions have strong correlation.[46,47] Staub et al reported that the primary site of metastatic cancer can be predicted based on the similarity between metastatic cancer and primary tissue.[46] 2) Some of the candidate genes were general tumor genes, such as PAGE4, a member of the GAGE family that is expressed in a variety of tumors.[41–43]

Genetic variations, such as single-nucleotide polymorphisms (SNPs) and copy number variations, have been proven to be a causal factor for tumorgenesis.[48–52] They can be used for cancer subtyping and drug response prediction.[22,48] Unfortunately, our dataset did not include genetic data. But based on central dogma and previous studies, most SNPs function through expression quantitative trait loci (eQTL).[17,18,53] The gene expression data can partially represent the effects of SNPs. If possible, we will preform DNA-Seq and RNA-Seq for the same patients and investigate the eQTL regulatory network of colorectal cancer patients with chemotherapy in the future.

Conclusion

Chemotherapy is a widely used treatment for cancers but not all cancer patients have expected responses to this treatment.

Table 5 The lowly expressed genes in FOLFOX responders

Gene name	Mean in NR[a]	Mean in R[b]	Fold change[c]
SLC26A9	81.0	31.5	0.39
ADAMTSL2	15.3	6.9	0.45
IGKC	2,452.8	1,261.2	0.51
TMPRSS3	298.0	175.0	0.59
CXorf57	79.4	46.9	0.59
OR10H2	13.3	8.2	0.62
HS3ST5	92.1	61.1	0.66

Notes: [a]NR, colorectal cancer patients without FOLFOX response. [b]R, colorectal cancer patients with FOLFOX response. [c]Fold change, R/NR.

In this study, we analyzed the gene expression profiles of FOLFOX responders and FOLFOX non-responders of colorectal cancer patients by combing several datasets. With advanced feature selection methods, we identified the biomarkers that can accurately predict the response of colorectal cancer patient to FOLFOX treatment. The biological analysis of selected genes revealed the possible mechanism of chemotherapy in colorectal cancer.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet*. 2014; 383(9927):1490–1502.
2. Abdolahi HM, Asiabar AS, Azami-Aghdash S, Pournaghi-Azar F, Rezapour A. Cost-effectiveness of Colorectal Cancer Screening and Treatment Methods: Mapping of Systematic Reviews. *Asia Pac J Oncol Nurs*. 2018;5(1):57–67.
3. Mohelnikova-Duchonova B, Melichar B, Soucek P. FOLFOX/FOLFIRI pharmacogenetics: the call for a personalized approach in colorectal cancer therapy. *World J Gastroenterol*. 2014;20(30):10316–10330.
4. Suh KW, Kim JH, Kim DY, Kim YB, Lee C, Choi S. Which gene is a dominant predictor of response during FOLFOX chemotherapy for the treatment of metastatic colorectal cancer, the MTHFR or XRCC1 gene? *Ann Surg Oncol*. 2006;13(11):1379–1385.
5. Kumar A, Peixoto RD, Kennecke HF, et al. Effect of Adjuvant FOLFOX Chemotherapy Duration on Outcomes of Patients With Stage III Colon Cancer. *Clin Colorectal Cancer*. 2015;14(4):262.e1–268.e1.
6. Watanabe T, Kobunai T, Yamamoto Y, et al. Gene expression signature and response to the use of leucovorin, fluorouracil and oxaliplatin in colorectal cancer patients. *Clin Transl Oncol*. 2011;13(6): 419–425.
7. Etienne-Grimaldi MC, Milano G, Maindrault-Goebel F, et al. Methylenetetrahydrofolate reductase (MTHFR) gene polymorphisms and FOLFOX response in colorectal cancer patients. *Br J Clin Pharmacol*. 2010;69(1):58–66.
8. Tsuji S, Midorikawa Y, Takahashi T, et al. Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis. *Br J Cancer*. 2012;106(1):126–132.
9. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy – analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004;20(3): 307–315.
10. Hubbell E, Liu WM, Mei R. Robust estimators for expression analysis. *Bioinformatics*. 2002;18(12):1585–1592.
11. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*. 2005;27(8):1226–1238.
12. Zhou Y, Zhang N, Li BQ, Huang T, Cai YD, Kong XY. A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis. *J Biomol Struct Dyn*. 2015;33(11): 2479–2490.
13. Zhao TH, Jiang M, Huang T, et al. A novel method of predicting protein disordered regions based on sequence features. *Biomed Res Int*. 2013; 2013:414327.
14. Niu B, Huang G, Zheng L, et al. Prediction of substrate-enzyme-product interaction based on molecular descriptors and physicochemical properties. *Biomed Res Int*. 2013;2013:674215.
15. Zhang N, Wang M, Zhang P, Huang T. Classification of cancers based on copy number variation landscapes. *Biochimica et Biophysica Acta (BBA) – General Subjects*. 2016;1860(11):2750–2755.

16. Liu L, Chen L, Zhang Y-H, et al. Analysis and prediction of drug–drug interaction by minimum redundancy maximum relevance and incremental feature selection. *J Biomol Struct Dyn*. 2017;35(2):312–329.
17. Li J, Huang T. Predicting and analyzing early wake-up associated gene expressions by integrating GWAS and eQTL studies. *Biochim Biophys Acta*. 2018;1864(6 Pt B):2241–2246.
18. Huang T, Cai Y-D. An Information-Theoretic Machine Learning Approach to Expression QTL Analysis. *PLoS One*. 2013;8(6):e67899.
19. Sun L, Yu Y, Huang T, et al. Associations between Ionomic Profile and Metabolic Abnormalities in Human Population. *PLoS One*. 2012; 7(6):e38845.
20. Zhang N, Huang T, Cai YD. Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. *Mol Genet Genomics*. 2015;290(1):343–352.
21. Shu Y, Zhang N, Kong X, Huang T, Cai Y-D. Predicting A-to-I RNA Editing by Feature Selection and Random Forest. *PLoS One*. 2014; 9(10):e110607.
22. Li BQ, You J, Huang T, Cai YD. Classification of non-small cell lung cancer based on copy number alterations. *PLoS One*. 2014;9(2): e88300.
23. Jiang Y, Huang T, Chen L, Gao Y-F, Cai Y, Chou K-C. Signal Propagation in Protein Interaction Network during Colorectal Cancer Progression. *Biomed Res Int*. 2013;2013(1):287019.
24. Zhang P-W, Chen L, Huang T, Zhang N, Kong X-Y, Cai Y-D. Classifying Ten Types of Major Cancers Based on Reverse Phase Protein Array Profiles. *PLoS One*. 2015;10(3):e0123147.
25. Huang T, Shu Y, Cai Y-D. Genetic differences among ethnic groups. *BMC Genomics*. 2015;16(1):1093.
26. Chen L, Li J, Zhang YH, et al. Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J Cell Biochem*. 2018;119(4):3394–3403.
27. Todorov H, Fournier D, Gerber S. Principal components analysis: theory and application to gene expression data analysis. *Genom Comput Biol*. 2018;4(2):e100041.
28. Szklo M, Nieto FJ. *Epidemiology Beyond the Basics*. Boston, MA: Jones and Bartlett; 2007.
29. Li J, Bi L, Shi Z, et al. RNA-Seq analysis of non-small cell lung cancer in female never-smokers reveals candidate cancer-associated long noncoding RNAs. *Pathol Res Pract*. 2016;212(6):549–554.
30. Yang H, Ma Y, Chen G, et al. Contribution of RIP3 and MLKL to immunogenic cell death signaling in cancer chemotherapy. *Oncoimmunology*. 2016;5(6):e1149673.
31. Sun D, Zhao L, Lin J, Zhao Y, Zheng Y. Cationic liposome co-encapsulation of SMAC mimetic and zVAD using a novel lipid bilayer fusion loaded with MLKL-pDNA for tumour inhibition in vivo. *J Drug Target*. 2018;26(1):45–54.
32. Fender AW, Nutter JM, Fitzgerald TL, Bertrand FE, Sigounas G. Notch-1 Promotes Stemness and Epithelial to Mesenchymal Transition in Colorectal Cancer. *J Cell Biochem*. 2015;116(11):2517–2527.
33. Vinson KE, George DC, Fender AW, Bertrand FE, Sigounas G. The Notch pathway in colorectal cancer. *Int J Cancer*. 2016;138(8):1835–1842.
34. Deshar R, Cho E-B, Yoon SK, Yoon J-B. CC2D1A and CC2D1B regulate degradation and signaling of EGFR and TLR4. *Biochem Biophys Res Commun*. 2016;480(2):280–287.
35. Nakamura A, Naito M, Tsuruo T, Fujita N. Freud-1/Aki1, a Novel PDK1-Interacting Protein, Functions as a Scaffold To Activate the PDK1/Akt Pathway in Epidermal Growth Factor Signaling. *Mol Cell Biol*. 2008;28(19):5996–6009.
36. Tuupanen S, Hänninen UA, Kondelin J, et al. Identification of 33 candidate oncogenes by screening for base-specific mutations. *Br J Cancer*. 2014;111(8):1657–1662.
37. Gerbe F, van Es JH, Makrini L, et al. Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *J Cell Biol*. 2011;192(5):767–780.
38. Notarnicola M, Messa C, Caruso MG. A significant role of lipogenic enzymes in colorectal cancer. *Anticancer Res*. 2012;32(7): 2585–2590.
39. Takasu S, Mutoh M, Takahashi M, Nakagama H. Lipoprotein Lipase as a Candidate Target for Cancer Prevention/Therapy. *Biochem Res Int*. 2012;2012(2):398697.
40. Vainio P, Gupta S, Ketola K, et al. Arachidonic Acid Pathway Members PLA2G7, HPGD, EPHX2, and CYP4F8 Identified as Putative Novel Therapeutic Targets in Prostate Cancer. *Am J Pathol*. 2011;178(2): 525–536.
41. Brinkmann U, Vasmatzis G, Lee B, Yerushalmi N, Essand M, Pastan I. PAGE-1, an X chromosome-linked GAGE-like gene that is expressed in normal and neoplastic prostate, testis, and uterus. *Proc Natl Acad Sci U S A*. 1998;95(18):10757–10762.
42. Iavarone C, Wolfgang C, Kumar V, et al. PAGE4 is a cytoplasmic protein that is expressed in normal prostate and in prostate cancers. *Mol Cancer Ther*. 2002;1(5):329–335.
43. Kong U, Koo J, Choi K, Park J, Chang H. The expression of GAGE gene can predict aggressive biologic behavior of intestinal type of stomach cancer. *Hepatogastroenterology*. 2004;51(59):1519–1523.
44. Chen Z, Li M, Yuan Y, Wang Q, Yan L, Gu J. Cancer/Testis Antigens and Clinical Risk Factors for Liver Metastasis of Colorectal Cancer: A Predictive Panel. *Dis Colon Rectum*. 2010;53(1):31–38.
45. Chen A-P, Chang M-H, Romero MF. Functional analysis of nonsynonymous single nucleotide polymorphisms in human SLC26A9. *Hum Mutat*. 2012;33(8):1275–1284.
46. Staub E, Buhr H-J, Gröne J. Predicting the site of origin of tumors by a gene expression signature derived from normal tissues. *Oncogene*. 2010;29(31):4485–4492.
47. Greco FA. Cancer of unknown primary site: still an entity, a biological mystery and a metastatic model. *Nat Rev Cancer*. 2014;14(1):3–4.
48. Huang T. *Copy Number Variations in Tumors*. Elsevier: Reference Module in Biomedical Sciences; 2018. Available from: https://doi.org/10.1016/B978-0-12-801238-3.65047-X. Accessed September 07, 2018.
49. Huang T, Li B-Q, Cai Y-D. The integrative network of gene expression, microRNA, methylation and copy number variation in colon and rectal cancer. *Curr Bioinform*. 2016;11(1):59–65.
50. Chen L, Huang T, Zhang Y-H, Jiang Y, Zheng M, Cai Y-D. Identification of novel candidate drivers connecting different dysfunctional levels for lung adenocarcinoma using protein-protein interactions and a shortest path approach. *Sci Rep*. 2016;6(1):29849.
51. Huang T, Yang J, Cai Y-D. Novel Candidate Key Drivers in the Integrative Network of Genes, MicroRNAs, Methylations, and Copy Number Variations in Squamous Cell Lung Carcinoma. *Biomed Res Int*. 2015;2015(2):358125.
52. Zhang TM, Huang T, Wang RF. Cross talk of chromosome instability, CpG island methylator phenotype and mismatch repair in colorectal cancer. *Oncol Lett*. 2018;16(2):1736–1746.
53. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*. 2015; 348(6235):648–660.

Supplementary materials

Table S1 The clinical information of the 144 colorectal cancer patients

Sample ID	Dataset	Actual response	Location	Gender	Predicted response
GSM1875899	GSE72970	Non-responder	Primary	Female	Responder
GSM1875907	GSE72970	Non-responder	Primary	Male	Non-responder
GSM1875917	GSE72970	Non-responder	Primary	Male	Responder
GSM1875935	GSE72970	Non-responder	Primary	Male	Responder
GSM1875937	GSE72970	Non-responder	Primary	Male	Responder
GSM1875938	GSE72970	Non-responder	Primary	Female	Responder
GSM1875947	GSE72970	Non-responder	Primary	Female	Non-responder
GSM1875952	GSE72970	Non-responder	Primary	Male	Responder
GSM1875959	GSE72970	Non-responder	Primary	Female	Non-responder
GSM1875989	GSE72970	Non-responder	Primary	Male	Non-responder
GSM1876008	GSE72970	Non-responder	Primary	Male	Non-responder
GSM1876009	GSE72970	Non-responder	Primary	Male	Non-responder
GSM1875897	GSE72970	Responder	Primary	Male	Responder
GSM1875898	GSE72970	Responder	Primary	Male	Responder
GSM1875900	GSE72970	Responder	Primary	Female	Responder
GSM1875902	GSE72970	Responder	Primary	Female	Responder
GSM1875914	GSE72970	Responder	Primary	Male	Responder
GSM1875916	GSE72970	Responder	Primary	Female	Responder
GSM1875918	GSE72970	Responder	Primary	Male	Responder
GSM1875919	GSE72970	Responder	Primary	Male	Non-responder
GSM1875920	GSE72970	Responder	Primary	Male	Responder
GSM1875923	GSE72970	Responder	Primary	Male	Responder
GSM1875924	GSE72970	Responder	Primary	Female	Responder
GSM1875929	GSE72970	Responder	Primary	Female	Responder
GSM1875932	GSE72970	Responder	Primary	Female	Responder
GSM1875948	GSE72970	Responder	Primary	Male	Responder
GSM1875954	GSE72970	Responder	Primary	Female	Non-responder
GSM1875955	GSE72970	Responder	Primary	Male	Responder
GSM1875956	GSE72970	Responder	Primary	Female	Responder
GSM1875969	GSE72970	Responder	Primary	Male	Responder
GSM1875972	GSE72970	Responder	Primary	Female	Responder
GSM1875981	GSE72970	Responder	Primary	Male	Responder
GSM710828	GSE28702	Non-responder	Metastasis	Female	Non-responder
GSM710829	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710830	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710831	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710832	GSE28702	Non-responder	Primary	Male	Responder
GSM710833	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710834	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710835	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710836	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710837	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710839	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710841	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710843	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710845	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710846	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710849	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710853	GSE28702	Non-responder	Metastasis	Female	Non-responder
GSM710855	GSE28702	Non-responder	Metastasis	Female	Non-responder
GSM710858	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710860	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710862	GSE28702	Non-responder	Primary	Female	Non-responder

(*Continued*)

Table S1 (*Continued*)

Sample ID	Dataset	Actual response	Location	Gender	Predicted response
GSM710863	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710865	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710867	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710869	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710871	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710873	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710905	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710906	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710908	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710911	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710913	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710915	GSE28702	Non-responder	Metastasis	Male	Non-responder
GSM710916	GSE28702	Non-responder	Metastasis	Female	Non-responder
GSM710918	GSE28702	Non-responder	Metastasis	Female	Non-responder
GSM710920	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710922	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710924	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710926	GSE28702	Non-responder	Primary	Female	Non-responder
GSM710928	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710930	GSE28702	Non-responder	Primary	Male	Non-responder
GSM710801	GSE28702	Responder	Metastasis	Female	Responder
GSM710802	GSE28702	Responder	Primary	Male	Non-responder
GSM710803	GSE28702	Responder	Primary	Male	Responder
GSM710804	GSE28702	Responder	Primary	Male	Non-responder
GSM710805	GSE28702	Responder	Primary	Male	Non-responder
GSM710806	GSE28702	Responder	Primary	Female	Responder
GSM710807	GSE28702	Responder	Primary	Male	Responder
GSM710808	GSE28702	Responder	Primary	Male	Responder
GSM710809	GSE28702	Responder	Primary	Male	Responder
GSM710810	GSE28702	Responder	Primary	Male	Responder
GSM710811	GSE28702	Responder	Primary	Male	Responder
GSM710812	GSE28702	Responder	Primary	Male	Responder
GSM710813	GSE28702	Responder	Metastasis	Male	Responder
GSM710814	GSE28702	Responder	Metastasis	Male	Responder
GSM710815	GSE28702	Responder	Metastasis	Male	Responder
GSM710816	GSE28702	Responder	Metastasis	Male	Responder
GSM710817	GSE28702	Responder	Metastasis	Male	Responder
GSM710818	GSE28702	Responder	Metastasis	Female	Responder
GSM710819	GSE28702	Responder	Metastasis	Male	Responder
GSM710820	GSE28702	Responder	Metastasis	Female	Responder
GSM710821	GSE28702	Responder	Primary	Male	Responder
GSM710822	GSE28702	Responder	Primary	Male	Responder
GSM710823	GSE28702	Responder	Primary	Male	Responder
GSM710824	GSE28702	Responder	Primary	Male	Responder
GSM710825	GSE28702	Responder	Primary	Female	Responder
GSM710826	GSE28702	Responder	Primary	Female	Responder
GSM710827	GSE28702	Responder	Primary	Female	Responder
GSM710875	GSE28702	Responder	Primary	Female	Responder
GSM710877	GSE28702	Responder	Primary	Male	Non-responder
GSM710879	GSE28702	Responder	Primary	Female	Non-responder
GSM710881	GSE28702	Responder	Metastasis	Female	Responder
GSM710883	GSE28702	Responder	Metastasis	Male	Responder
GSM710885	GSE28702	Responder	Primary	Male	Responder
GSM710886	GSE28702	Responder	Primary	Male	Responder
GSM710888	GSE28702	Responder	Primary	Male	Responder
GSM710890	GSE28702	Responder	Primary	Male	Responder

(*Continued*)

Table S1 (*Continued*)

Sample ID	Dataset	Actual response	Location	Gender	Predicted response
GSM710892	GSE28702	Responder	Primary	Female	Non-Responder
GSM710894	GSE28702	Responder	Primary	Female	Responder
GSM710896	GSE28702	Responder	Primary	Female	Responder
GSM710898	GSE28702	Responder	Primary	Female	Responder
GSM710900	GSE28702	Responder	Primary	Male	Responder
GSM710902	GSE28702	Responder	Primary	Female	Responder
GSM496015	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496016	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496017	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496018	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496019	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496022	GSE19860	Non-responder	Primary	NA	Responder
GSM496023	GSE19860	Non-responder	Primary	NA	Responder
GSM496024	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496025	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496026	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496028	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496029	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496031	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496032	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496033	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496034	GSE19860	Non-responder	Primary	NA	Responder
GSM496035	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496037	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496038	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496042	GSE19860	Non-responder	Primary	NA	Non-responder
GSM496020	GSE19860	Responder	Primary	NA	Responder
GSM496021	GSE19860	Responder	Primary	NA	Non-responder
GSM496027	GSE19860	Responder	Primary	NA	Responder
GSM496030	GSE19860	Responder	Primary	NA	Responder
GSM496036	GSE19860	Responder	Primary	NA	Responder
GSM496039	GSE19860	Responder	Primary	NA	Responder
GSM496040	GSE19860	Responder	Primary	NA	Responder
GSM496041	GSE19860	Responder	Primary	NA	Non-responder
GSM496043	GSE19860	Responder	Primary	NA	Non-responder

Abbreviation: NA, not applicable

Table S2 The top 138 mRMR genes

Order	Name	Score
1	*LOC100009676*	0.131
2	*ZNF461*	0.101
3	*MLKL*	0.072
4	*MGC15885*	0.083
5	*MBTD1*	0.071
6	*CC2D1A*	0.067
7	*FAM104A*	0.061
8	*KIF3B*	0.06
9	*SYTL1*	0.06
10	*EML6*	0.057
11	*ENSG00000244627*	0.057
12	*AHCYL1*	0.058
13	*OR10H2*	0.057
14	*CYP4F8*	0.058
15	*LTA4H*	0.055
16	*JOSD2*	0.055
17	*FAM120C*	0.055
18	*IQSEC2*	0.053
19	*C11orf9*	0.053
20	*CRYBB1*	0.051
21	*SLC16A4*	0.052
22	*TBC1D21*	0.051
23	*TMEM160*	0.05
24	*NIP7*	0.05
25	*ULBP1*	0.05
26	*C15orf26*	0.049
27	*ATP6V1B2*	0.048
28	*DRAP1*	0.047
29	*C12orf34*	0.047
30	*LHX9*	0.047
31	*NPEPPS*	0.046
32	*ZNF569*	0.046
33	*LPL*	0.045
34	*ENSG00000240024*	0.044
35	*P2RX4*	0.044
36	*GSTM3*	0.043
37	*FOSL2*	0.043
38	*PDK4*	0.042
39	*COX8A*	0.042
40	*NR4A2*	0.042
41	*BPTF*	0.042
42	*LIPF*	0.04
43	*HAUS1*	0.04
44	*SLC17A7*	0.04
45	*PRR14*	0.04
46	*PDE10A*	0.04
47	*SUPT4H1*	0.039
48	*PIGW*	0.039
49	*TM4SF5*	0.039
50	*PECR*	0.039
51	*COMT*	0.039
52	*IGKC*	0.039
53	*MOBKL3*	0.038
54	*NOL6*	0.038
55	*REG3G*	0.038
56	*TMEM66*	0.037
57	*PATE2*	0.036
58	*JUND*	0.037
59	*IL17D*	0.036
60	*ENSG00000186056*	0.036
61	*ADAMTSL2*	0.036
62	*TMPRSS3*	0.035
63	*ATP10A*	0.036
64	*GRK4*	0.036
65	*NEUROG3*	0.035
66	*WASF2*	0.035
67	*HIAT1*	0.035
68	*NFIA*	0.035
69	*LOC284100*	0.034
70	*IFT81*	0.034
71	*GSTZ1*	0.034
72	*ENSG00000235471*	0.034
73	*CXorf57*	0.034
74	*OXCT2*	0.034
75	*LRRC55*	0.034
76	*DHX58*	0.034
77	*RNF25*	0.034
78	*SLC26A9*	0.034
79	*ZNF140*	0.033
80	*ENSG00000231078*	0.033
81	*HOXA11AS*	0.032
82	*SEC14L2*	0.032
83	*IQCG*	0.032
84	*CACNG4*	0.032
85	*DDX42*	0.032
86	*C14orf102*	0.032
87	*HSPA4*	0.032
88	*INTS10*	0.032
89	*ENSG00000229522*	0.031
90	*ZHX3*	0.031
91	*LOC100130155*	0.031
92	*LOC284648*	0.031
93	*BDKRB2*	0.031
94	*NCRNA00116*	0.031
95	*HDLBP*	0.031
96	*KRT74*	0.031
97	*ZNF528*	0.03
98	*SPG7*	0.03
99	*MORF4L1*	0.03
100	*LOC340107*	0.03
101	*DNAJC5G*	0.03
102	*C16orf92*	0.03
103	*ZNF204P*	0.029
104	*DNAJC2*	0.029
105	*RBKS*	0.029
106	*PACSIN1*	0.029
107	*ANKMY1*	0.029
108	*NCRNA00173*	0.029
109	*ZNF205*	0.029
110	*PPP1R1C*	0.029
111	*FUT4*	0.029
112	*ZNF605*	0.029
113	*RNF187*	0.028
114	*RUNDC1*	0.028

Table S2 (*Continued*)

Order	Name	Score
115	*COX4NB*	0.028
116	*TNFRSF1A*	0.028
117	*IRF3*	0.028
118	*HS3ST5*	0.028
119	*POM121*	0.028
120	*VIT*	0.028
121	*NPEPL1*	0.028
122	*DMC1*	0.028
123	*ATP13A2*	0.028
124	*C20orf194*	0.028
125	*TTC21B*	0.028
126	*EIF4B*	0.027
127	*PAGE4*	0.027
128	*SOCS6*	0.027
129	*MNAT1*	0.027
130	*LMOD3*	0.027
131	*ABCD4*	0.027
132	*MTMR4*	0.027
133	*HMGCL*	0.027
134	*ZNHIT3*	0.027
135	*CD151*	0.027
136	*SEP15*	0.026
137	*SRXN1*	0.026
138	*NDUFA8*	0.026